Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance

Year: 2020 | Citations: 713 | Authors: Gagan Bansal, Tongshuang Sherry Wu, Joyce Zhou, Raymond Fok, Besmira Nushi

---

Abstract

Many researchers motivate explainable AI with studies showing that human-AI team performance on decision-making tasks improves when the AI explains its recommendations. However, prior studies observed improvements from explanations only when the AI, alone, outperformed both the human and the best team. Can explanations help lead to complementary performance, where team accuracy is higher than either the human or the AI working solo? We conduct mixed-method user studies on three datasets, where an AI with accuracy comparable to humans helps participants solve a task (explaining itself in some conditions). While we observed complementary improvements from AI augmentation, they were not increased by explanations. Rather, explanations increased the chance that humans will accept the AI's recommendation, regardless of its correctness. Our result poses new challenges for human-centered AI: Can we develop explanatory approaches that encourage appropriate trust in AI, and therefore help generate (or improve) complementary performance?